BIOMETRICS 46, 963–975
December 1990

The Design and Analysis of Case–Control Studies with Biased Sampling

Clarice R. Weinberg

Division of Biometry and Risk Assessment,
National Institute of Environmental Health Sciences,
P.O. Box 12233, Research Triangle Park, North Carolina 27709, U.S.A.

and

Sholom Wacholder

National Cancer Institute, Biostatistics Branch, EPN Room 403,
Bethesda, Maryland 20892, U.S.A.

SUMMARY

A design is proposed for case–control studies in which selection of subjects for full variable ascertainment is based jointly on disease status and on easily obtained "screening" variables that may be related to the disease. Recruitment of subjects follows an independent Bernoulli sampling scheme, with recruitment probabilities set by the investigator in advance. In particular, the sampling can be set up to achieve, on average, frequency matching, provided prior estimates of the disease rates or odds ratios associated with screening variables such as age and sex are available. Alternatively—for example, when studying a rare exposure—one can enrich the sample with certain categories of subject. Following such a design, there are two valid approaches to logistic regression analysis, both of which allow for efficient estimation of effects associated with the screening variables that were allowed to bias the recruitment. The statistical properties of the estimators are compared, both for large samples, based on asymptotics, and for small samples, based on simulations.

1. Introduction

Matching strategies are often employed to improve efficiency in case–control studies. For example, controls can be chosen to have the same empirical distribution as cases for some easily obtained variables known to be related to disease risk, such as age and sex. This technique, known as "frequency" or "quota" matching, imposes balance but introduces both practical and statistical complications. If recruitment is ongoing, with cases recruited as they are diagnosed, there is no way to know how many controls in a particular stratum (defined by the matching factors) will ultimately be required. Thus there are unavoidable recruitment inefficiencies. In the analysis, the investigator must incorporate the matching factors into the risk model (Gail, 1988), but must remember that the resulting point estimates for effects associated with those factors are meaningless (Kleinbaum, Kupper, and Morgenstern, 1982, p. 382). Furthermore, since the main effects for matching variables cannot be estimated, additive models for interactions involving those variables cannot be assessed. Recently Thomas and Greenland (1985) have summarized these problems, noting that "any potential gain in statistical efficiency derived from matching must . . . be weighed against several disadvantages: (a) after matching, the main effect of a variable cannot be

Key words: Biased sampling; Case–control studies; Matching; Randomized recruitment; Two-stage sampling.

tested or estimated; (b) matching precludes fitting nonmultiplicative models; (c) each additional matching factor can add cost and complexity to the sampling scheme and increase the risk of being unable to find a match."

This paper proposes a design strategy which, when there is some prior information about the risk associated with the matching factors, can solve the first two of these problems and greatly alleviate the third. In the proposed method, each potential subject identified is invited to participate or not according to a Bernoulli random mechanism, with probabilities set by the investigator. The recruitment probabilities can depend jointly on disease status and on any easily ascertained screening variables, such as sex and age. Controls are recruited independently of one another and of cases, without any quota requirements imposed by already recruited study subjects. This design is particularly advantageous in a situation where the exposure of interest is expensive to measure and the screening variables are known risk factors that are relatively cheap to ascertain.

Following such a design, there are two valid choices for analysis. One can condition on the numbers of screened and recruited individuals and carry out a conditional maximum pseudo-likelihood analysis, resembling that proposed by Breslow and Cain (1988) for "two-stage" sampling. Alternatively, individuals who were screened but not recruited can be ignored, so that the stochastic recruitment process is absorbed into the aggregate of random processes giving rise to the sample. With incorporation of proper "offsets," based on the recruitment probabilities, logistic regression analysis by maximum likelihood [e.g., by GLIM (Baker and Nelder, 1978)] goes through without modification, and all coefficients and variances can be consistently estimated, including those associated with factors that were allowed to bias the recruitment. Exploratory analyses then proceed with great flexibility, because likelihood ratio testing is valid for model comparisons.

Section 2 considers, as an example, the problem of studying household exposure to radon as an etiologic factor in lung cancer, where current smoking status is an important screening variable. A method is developed and illustrated for computing recruitment probabilities needed to achieve a specified distribution of screening variables, when the case and population distributions are approximately known. Section 3 sets out the notation for randomized recruitment in the context of a logistic model, and describes the incorporation of the fixed and known correction parameters, for both unconditional and conditional maximum likelihood analysis. A generalization to other risk models is also provided. Section 4 adapts the method of Breslow and Cain to this setting and gives asymptotic relative efficiencies for the log odds ratio for two factors following stochastic frequency matching, for selected combinations of parameters. Section 5 compares the maximum partial likelihood approach to the conditional approach adapted from Breslow and Cain, describing the results of a small simulation study. Section 6 discusses advantages and disadvantages of randomized recruitment, and offers guidelines for choice of analysis.

2. The Design, and an Example

Consider, as a particular example, the problem of studying household exposure to radon gas and its possible etiologic link to lung cancer. Estimates of the number of deaths in the United States per year due to environmental radon exposure vary from about 9,000 to 25,000 (Cohen, 1987). These estimates of its public health impact are imprecise, based as they are on data from miner cohorts extrapolated down to the relatively low levels experienced in the typical home. Extensive direct epidemiologic evidence related to effects of low-level radon exposure does not yet exist, but studies are in progress or in planning stages. The focus of such a study might be to estimate effects of radon and also to assess the joint effects of radon exposure and cigarette smoking.

There is a fundamental problem in trying to identify etiologic factors for lung cancer: because most lung cancer is caused by smoking, a random series of cases is likely to include only a small proportion of nonsmokers, perhaps 15%, while a random sample of controls (in the same age group) may include about 70% nonsmokers. Stratification on smoking status is clearly required, but will result in gross imbalance: cases will be plentiful in the smokers' stratum but scarce in the nonsmokers' stratum. If the odds ratio associated with lifetime radon exposure, dichotomized at 4 pCi/L, is 1.5, and if 10% of controls have that level of exposure, then under random sampling about 1,000 cases and 1,000 controls are required to achieve 80% power to detect the radon effect at a .05 significance level (see Woolson, Bean, and Rojas, 1986). By contrast, if the study can be designed so that equal numbers of cases and controls are smokers, then 725 of each would be required to achieve the same power. The extent of the gain here in efficiency is attributable to the very large odds ratio associated with smoking; matching on a weaker risk factor would bring a less impressive efficiency advantage.

Estimating lifetime radon exposure by making measurements on all current and former residences for each person studied is a time-consuming and expensive process, while determining smoking information is relatively easy. There are clear advantages to a design that effectively matches cases and controls on smoking status.

2.1 *Stochastic Frequency Matching*

The investigator can achieve approximate frequency matching by selecting all cases identified, and selecting controls according to a biased sampling scheme. Each potential control is first randomly identified and screened. A second random selection will govern whether a person is actually recruited, i.e., invited to participate in the full study. Suppose that the population distribution for a K-level screening factor is estimated to be $(c_1, c_2, c_3, \ldots, c_K)$ for cases and $(h_1, h_2, h_3, \ldots, h_K)$ for nondiseased persons. Let r_{ij} denote the recruitment probability for an individual screened with disease status i ($i = 1$ for cases and $i = 0$ for controls) and level j of the screening variable. Set $r_{1j} = 1$ for all j and set

$$r_{0j} = \frac{c_j/h_j}{\max(c_i/h_i)}.$$

Sampling of controls is to be done according to a Bernoulli sampling mechanism, where for each potential control screened a random uniform (0, 1) number is generated. If the result is less than the above number, the person is invited to participate in the full study. Then, conditional on the total number of controls recruited, the sampling distribution of the controls becomes a multinomial with probabilities $(c_1, c_2, c_3, \ldots, c_K)$, the same as that of cases. (In practice this equality of case and control distribution will only be approximate, since the case distribution must be estimated.) This sampling scheme minimizes the number of controls that must be screened, since the most oversampled control category is sampled with probability 1. This scheme will be called "stochastic frequency matching."

As an example of its application assume that the distributions of smoking habits among cases and controls are known to be approximately as in Table 1. To achieve approximate balance without discarding any recruitable cases, one would recruit available cases with probability 1 and nonsmoking controls with probability (.15/.7)/(.42/.1) = .051. Nondiseased light smokers would be recruited with probability .357, moderate smokers with probability .6667, and heavy smokers with probability 1.0. The result would not be perfect frequency matching balance, but balance in expectation. Since only about 24% of controls (the weighted average of the recruitment probabilities) would be recruited, approximately

Table 1
Hypothetical distribution of smoking status among lung cancer cases and controls

	Nonsmokers	Light	Moderate	Heavy
Cases	.15	.15	.28	.42
Controls	.70	.10	.10	.10

4 times as many controls as cases would need to be screened in order to recruit equal numbers (in expectation) of cases and controls.

Suppose the investigator's initial estimates for the distributions of the matching factors were wrong and therefore the "wrong" recruitment probabilities were used. Both analyses to be described below will still be valid, since it will nowhere be assumed that the sampling probabilities were derived from true distributions.

Now suppose the investigator has no prior knowledge of the distributions of the screening factors, but does have access either to disease rates or to relative risk estimates for the disease under study, based on the screening variables only. For example, we might have a good risk model for lung cancer as related to age, sex, and smoking data, but not have good information on population smoking patterns for, say, Utah. Recruitment probabilities yielding stochastic frequency matching can still be set up. Suppose the odds ratio for lung cancer associated with stratum j is estimated to be OR_j. Let r_{1j} be 1, as before, for all levels of the screening variables, and let r_{0j} be $\mathrm{OR}_j/(\max \mathrm{OR}_i)$ for all j. This will again achieve approximate balance between cases and controls in their distributions of screening factors. To see why this works observe that if π_j is the prevalence of category j among nondiseased people, then the probability that a control sampled is in category j $(\pi_j \mathrm{OR}_j/(\max \mathrm{OR}_i))$ divided by the corresponding probability for category 1 $(\pi_1/(\max \mathrm{OR}_i))$, matches that of cases.

2.2 *General Target Distributions*

An investigator may wish to select recruitment fractions to achieve a specified target distribution, $(t_1, t_2, t_3, \ldots, t_K)$, where the population distribution is again $(c_1, c_2, c_3, \ldots, c_K)$ for diseased and $(h_1, h_2, h_3, \ldots, h_K)$ for nondiseased individuals. For example, results in Breslow and Cain (1988, Table 3a) suggest that for assessment of interaction, when the exposure is rare (the exposure being known here at the screening stage), it is highly advantageous to equate the numbers of exposed and unexposed individuals in each disease category. To achieve any specified target distribution, while minimizing the number who must be screened, set the r_{1j} for each case identified to be

$$\frac{t_i/c_i}{\max(t_j/c_j)}$$

and an analogous expression for controls.

If the sampling is done according to an independent Bernoulli mechanism for each potential subject identified, then the distribution for recruited cases and for recruited controls will still be multinomial, but now with parameters $(t_1, t_2, t_3, \ldots, t_K)$, as if (in the distorted universe in which we are sampling) this is the true common distribution. A standard classical logistic regression analysis based on the recruited individuals from the population represented in Table 1 can then be carried out but will wrongly estimate an odds ratio of 1 for each smoking category, compared to nonsmokers.

There may also be situations where different target distributions are desired for cases and controls. For example, because of time and cost considerations one might be willing to

tolerate an excess of smokers among the cases. Distinct target distributions can be achieved by a straightforward generalization of the method described above.

Note that the Bernoulli subsampling is not just a matter of convenience. One might think, for example, that sampling every third control in a particular category would be equivalent (at least asymptotically) to applying Bernoulli subsampling with probability $\frac{1}{3}$. To see why this is not so, recall that if Y is binomially distributed then αY is not, unless α is 0 or 1. Whether an individual is included or not must be independent of which other individuals are included. Only Bernoulli sampling in the proposed design allows us to distort (within each disease category) the probability distributions for the screening variables, while preserving independence among subjects.

2.3 *Sample Enrichment*

In some situations one might wish to oversample among certain categories that would come up only rarely with random sampling. For example, in a study of radon and lung cancer one might wish to oversample nonsmoking cases to enhance our ability to characterize the radon/smoking interaction by providing closer to equal numbers of smokers and nonsmokers. This oversampling can be done without any prior knowledge of the distribution of the factor. For example, a rare exposure suspected of being a risk factor (but available as a screening variable) may be oversampled among cases, and even more strongly oversampled among controls. For example, one could sample exposed persons with probability 1, unexposed cases with probability .5, and unexposed controls with probability .2. It is easily shown that this strategy would be particularly advantageous (leading to balance) under a scenario where the true relative risk is 2.5.

3. Maximum Partial Likelihood Logistic Analysis

We next describe methodology for analyzing data arising from a design that uses biased sampling, and demonstrate that all effects, including those associated with screening factors allowed to bias the recruitment, can be estimated consistently.

3.1 *Notation and Model Specification*

Suppose $D = 1$ or 0 according as the disease is present or not, and $R = 1$ or 0 according as the person is recruited into the study or not. In order to be recruited for full participation in the study, the potential subject must first be identified. It is conceptually useful to partition sampling into two events. First there is an identification event, denoted by $I = 1$, that does depend on disease status but does *not* depend on other variables. In practice the investigator typically has one sampling system (e.g., a cancer registry) that samples randomly within available cases and another system (e.g., random digit dialing) that samples randomly within potential controls from the same population. The second step is where the subject is invited to participate in the full study. In this second step the investigator may allow the probability of final recruitment to depend jointly on both disease status and on any screening variables. Suppose disease risk in the population can be characterized by a vector, $\mathbf{X}$, of variables, including the exposure of interest and interactions, with a 1 entered as the first component, to allow estimation of the mean parameter. Assume that the logistic model is appropriate for disease risk in the population, so that

$$\ln\left(\frac{\Pr[D = 1 \mid \mathbf{X}]}{\Pr[D = 0 \mid \mathbf{X}]}\right) = \mathbf{X}\beta$$

for some column vector, β, of regression coefficients.

Note the following factorization:

$$\Pr[R = 1, I = 1, D = i \mid \mathbf{X}] = \Pr[D = i \mid \mathbf{X}]\Pr[R = 1, I = 1 \mid D = i, \mathbf{X}].$$

It follows from this that whatever sampling strategy has been followed, the biased study logit can be separated into two terms, as follows:

$$\begin{aligned}
&\ln\left(\frac{\Pr[D = 1 \mid R = 1, I = 1, \mathbf{X}]}{\Pr[D = 0 \mid R = 1, I = 1, \mathbf{X}]}\right) \\
&\quad = \ln\left(\frac{\Pr[R = 1, I = 1, D = 1 \mid \mathbf{X}]}{\Pr[R = 1, I = 1, D = 0 \mid \mathbf{X}]}\right) \\
&\quad = \ln\left(\frac{\Pr[D = 1 \mid \mathbf{X}]\Pr[R = 1, I = 1 \mid D = 1, \mathbf{X}]}{\Pr[D = 0 \mid \mathbf{X}]\Pr[R = 1, I = 1 \mid D = 0, \mathbf{X}]}\right) \\
&\quad = \ln\left(\frac{\Pr[D = 1 \mid \mathbf{X}]}{\Pr[D = 0 \mid \mathbf{X}]}\right) + \ln\left(\frac{\Pr[R = 1, I = 1 \mid D = 1, \mathbf{X}]}{\Pr[R = 1, I = 1 \mid D = 0, \mathbf{X}]}\right), \qquad (1)
\end{aligned}$$

where the first term is the logit based on the true population and the second is the error related to sampling bias.

3.2 *Unconditional Logistic Regression*

Since we assume the initial identification step depends on disease status but not on covariates, we can factor as follows:

$$\begin{aligned}
\Pr[R = 1, I = 1 \mid D = i, \mathbf{X}] &= \Pr[I = 1 \mid D = i, \mathbf{X}]\Pr[R = 1 \mid I = 1, D = i, \mathbf{X}] \\
&= \Pr[I = 1 \mid D = i]\Pr[R = 1 \mid I = 1, D = i, \mathbf{X}].
\end{aligned}$$

It follows from this that the second term in (1) can be rewritten as follows:

$$\begin{aligned}
&\ln\left(\frac{\Pr[R = 1, I = 1 \mid D = 1, \mathbf{X}]}{\Pr[R = 1, I = 1 \mid D = 0, \mathbf{X}]}\right) \\
&\quad = \{\ln(\Pr[I = 1 \mid D = 1]) - \ln(\Pr[I = 1 \mid D = 0])\} \\
&\qquad + \{\ln(\Pr[R = 1 \mid I = 1, D = 1, \mathbf{X}]) - \ln(\Pr[R = 1 \mid I = 1, D = 0, \mathbf{X}])\}.
\end{aligned}$$

Note that the first difference term is constant, and the second difference term, the part involving recruitment, is within the control of the investigator, and is to be fixed as part of the design. Now if we make the very mildly restrictive assumption that this can be specified additively, then we can write, for some known column vector, $\boldsymbol{\gamma}$:

$$\mathbf{X}\boldsymbol{\gamma} = \ln(\Pr[R = 1 \mid I = 1, D = 1, \mathbf{X}]) - \ln(\Pr[R = 1 \mid I = 1, D = 0, \mathbf{X}]).$$

If we let $\boldsymbol{\beta}^*$ denote the limit of the logistic regression maximum likelihood estimator, conditional on recruitment, notice that

$$\text{logit}(\Pr[D = 1 \mid R = 1, \mathbf{X}]) = \mathbf{X}\boldsymbol{\beta}^* = \mathbf{X}\boldsymbol{\beta} + \mathbf{X}\boldsymbol{\gamma},$$

and it follows that if $\hat{\boldsymbol{\beta}}^*$ is the maximum likelihood estimator based on a "prospective" analysis (as developed by Prentice and Pyke, 1979), then $\hat{\boldsymbol{\beta}}^* - \boldsymbol{\gamma}$ is the maximum likelihood estimator for $\boldsymbol{\beta}$. Furthermore, since the correction, $\mathbf{X}\boldsymbol{\gamma}$, is a constant "offset" in the model, the variance–covariance is correctly estimated with no correction required. Maximum likelihood estimation is straightforward in GLIM (Baker and Nelder, 1978), where one simply specifies $\mathbf{X}\boldsymbol{\gamma}$ as an "offset."

More generally, the bias need not be expressible as a linear function of the covariates. We can compute maximum likelihood estimates using GLIM by declaring for each individual the known error, $\ln(\Pr[R = 1 \mid I = 1, D = 1, \mathbf{X}] - \ln(\Pr[R = 1 \mid I = 1, D = 0, \mathbf{X}])$ as an offset. In fact, the screening process (hence the offset) may even depend on variables not included in the final model. Since this maximum likelihood computation uses the offsets but ignores the screening information for subjects not recruited for full participation in the study, it will be referred to in what follows as maximum partial likelihood estimation, or MPLE. The Appendix develops an analogous result for conditional logistic regression.

As an example, considering the population of Table 1, let s_1, s_2, s_3, and s_4 denote indicator variables for the four levels of smoking status. Then the offset for the maximum partial likelihood approach would be $\sum s_j \ln(r_{1j}/r_{0j})$. Thus for stochastic frequency matching the offsets corresponding to the four smoking categories would be $-\ln(.051)$ for nonsmokers, $-\ln(.357)$ for light smokers, $-\ln(.66667)$ for moderate smokers, and 0.0 for heavy smokers.

3.3 *More General Risk Models*

As suggested in the Introduction, the proposed design allows for fitting nonmultiplicative models for describing interaction. Suppose the risk $\Pr[D = 1 \mid \mathbf{X}]$ is specified by some function $f(\mathbf{X})$. Then one can show that $\text{logit}(\Pr[D = 1 \mid R = 1, I = 1, \mathbf{X}]) = K + S(\mathbf{X}) + \text{logit}[f(\mathbf{X})]$, where $S(\mathbf{X})$ is the same log ratio of sampling probabilities used as an offset above and K is the log ratio of probabilities of identification for cases and controls. Thus nonmultiplicative models can also be fit by maximum likelihood following this design, although in general this will require additional software development.

4. Generalization of the Method of Breslow and Cain

An alternative approach is suggested by the "two-stage" design proposed by White (1982); see also Walker (1982). Briefly, White considered the situation where exposure information is already available for a large sample of cases and controls at stage 1 (screening). Complete covariable ascertainment is then carried out only on a subsample, where sampling fractions can depend jointly on disease status and covariates. White showed how to compute a valid estimate of the adjusted exposure odds ratio, by incorporating information from the initial, complete sample. Recently, Breslow and Cain (1988) extended this approach by allowing for a multilevel exposure variable, and any number and type of covariables.

The technique described by Breslow and Cain can also be applied following ongoing randomized recruitment, if we simply condition on the actual numbers screened and recruited in the various disease/variable subgroups. Conceptually, the "first-stage" sample becomes all those who were screened and judged eligible for randomization (for whom disease status and screening information were available), and the "second-stage" sample is all those who were then, based on that partial information and a Bernoulli sampling mechanism, randomized to recruitment and complete variable ascertainment.

Returning to the example of Table 1, let s_1, s_2, s_3, and s_4 again denote indicator variables for the four levels of smoking status. Suppose that following individual randomization a total of n_{ij} out of N_{ij} were randomized to recruitment. Then the offset for the maximum partial likelihood approach of Section 3 would be $\sum s_j \ln(r_{1j}/r_{0j})$, while that corresponding to the maximum conditional pseudo-likelihood (as developed by Breslow and Cain) would be $\sum s_j \ln(n_{1j}N_{0j}/(n_{0j}N_{1j}))$. The variance estimate for the maximum partial likelihood approach is then the naive variance, obtainable for example from GLIM. The variance estimator for the maximum pseudo-likelihood approach is given by Breslow and Cain.

It is worth noting that if the investigator has used one form of a variable in setting up the recruitment probabilities and then wishes to use a more detailed form of the variable in the analysis, this can be done quite easily (Cain and Breslow, 1988). For example, in the radon study discussed above, one might wish to model the dose response for smoking using more detailed second-stage information, e.g., number of cigarettes smoked per day, rather than on the crude categorization given in Table 1. To do this, either with the method of Breslow and Cain or with the maximum partial likelihood estimation of Section 3, declare the appropriate offsets based on the probabilities actually used in recruitment, but omit the screening form of the variable from the model.

5. Comparison of Maximum Partial Likelihood with the Method of Breslow and Cain

Just how efficient is stochastic frequency matching under these methods of analysis? Consider the simple situation where there are two dichotomous factors and recruitment is biased so that cases and controls are, in expectation, matched along factor 1. Suppose, as in Breslow and Cain, that factor 1 occurs with prevalence .05 and factor 2 with prevalence .30 in the nondiseased population. Here factor 1 is considered to be a screening variable. The asymptotic relative efficiencies depend on the odds ratio relating the two factors in the population, to be denoted θ. Table 2 lists asymptotic relative efficiencies for the estimated log odds ratios, β_1 and β_2, for the two factors using the method of Breslow and Cain, and also for the maximum partial likelihood method (MPLE). For both methods of analysis stochastic frequency matching is compared to a design where equal numbers of cases and controls are studied but the sampling is purely random within disease category. The sampling for both designs retains all cases and a proportion of controls, so that the "first-stage" sample sizes are the same. The difference is that under random sampling the second-stage recruitment is done without regard to factor 1 status.

Note from Table 2 that the effect of the matching factor not only can be estimated under stochastic frequency matching, but can often be estimated with better precision than under random sampling. This precision advantage is most apparent under the MPLE method (comparing columns 4 and 5). This is only because the method of Breslow and Cain makes better use of the first-stage data under the random sampling design. Column 6 lists the relative efficiencies of MPLE and the Breslow/Cain method when both are applied following stochastic frequency matching, and shows that the Breslow/Cain method is consistently slightly more efficient for estimating β_1.

The precision of estimation for β_2 is also enhanced (here both methods give the same variance), as we would expect. The rightmost column shows the asymptotic relative efficiency associated with the interaction parameter, β_3, under a model where the true interaction is 0. The advantage to this design appears to be strongest for the estimation of interaction. Results given by Breslow and Cain (1988, Table 3a) suggest that even greater gains in efficiency for assessing interaction are possible by using as the target distribution (as in Section 2.2) half with and half without the exposure (or screening factor), rather than stochastic frequency matching.

The variances associated with the maximum partial likelihood procedure are identical to those following the method of Breslow and Cain, except for estimation of the effects involving only screening variables (hence only one column is shown for asymptotic relative efficiencies for β_2 and β_3). Thus if our primary objective is to estimate effects associated with a second-stage variable, there is no efficiency advantage to the method of Breslow and Cain. Notice from Table 2 that the maximum partial likelihood approach to estimating the odds ratio associated with the screening variable typically did only slightly worse than the method of Breslow and Cain.

Maximum partial likelihood is always less efficient than the method of Breslow and Cain for estimating effects of first-stage variables (here β_1), since (see their Proposition 3) the

Table 2
Asymptotic relative efficiency for stochastic frequency matching relative to random case–control sampling

			β_1				
$\exp(\beta_1)$	$\exp(\beta_2)$	θ	BC	MPLE	BC/MPLE	β_2	β_3
2.0	.2	.2	1.01	1.57	.97	.99	1.24
2.0	.2	.5	1.03	1.50	.97	1.00	1.23
2.0	.2	1.0	1.04	1.41	.98	1.01	1.21
2.0	.2	2.0	1.04	1.29	.99	1.01	1.18
2.0	.2	5.0	1.02	1.10	1.00	1.01	1.08
2.0	1.0	.2	1.00	1.40	.98	.99	1.43
2.0	1.0	.5	1.00	1.40	.98	1.00	1.42
2.0	1.0	1.0	1.00	1.41	.98	1.01	1.41
2.0	1.0	2.0	1.00	1.40	.98	1.01	1.40
2.0	1.0	5.0	1.00	1.38	.98	1.01	1.40
2.0	5.0	.2	1.00	1.08	1.00	1.00	1.12
2.0	5.0	.5	1.02	1.22	.99	1.00	1.30
2.0	5.0	1.0	1.04	1.39	.98	1.01	1.41
2.0	5.0	2.0	1.05	1.57	.97	1.00	1.48
2.0	5.0	5.0	1.04	1.75	.95	.98	1.50
10.0	.2	.2	1.00	2.85	.83	.99	2.09
10.0	.2	.5	1.05	2.87	.84	1.08	2.01
10.0	.2	1.0	1.13	2.84	.86	1.18	1.97
10.0	.2	2.0	1.23	2.72	.88	1.27	1.97
10.0	.2	5.0	1.28	2.35	.92	1.29	2.07
10.0	1.0	.2	1.00	2.75	.86	1.15	3.46
10.0	1.0	.5	1.00	2.84	.85	1.07	3.11
10.0	1.0	1.0	1.00	2.88	.84	1.16	2.88
10.0	1.0	2.0	1.01	2.84	.85	1.21	2.75
10.0	1.0	5.0	1.03	2.65	.87	1.19	2.80
10.0	5.0	.2	1.06	2.19	.92	1.01	3.35
10.0	5.0	.5	1.10	2.49	.89	1.10	3.17
10.0	5.0	1.0	1.11	2.70	.87	1.16	2.90
10.0	5.0	2.0	1.05	2.73	.84	1.15	2.63
10.0	5.0	5.0	.93	2.53	.83	1.04	2.45

BC denotes the ratios (for the two designs) of asymptotic variances computed using the method of Breslow and Cain (1988), and MPLE denotes the ratios based on maximum partial likelihood estimation. The column labelled BC/MPLE shows the variance under BC divided by that under MPLE, following stochastic frequency matching. As in Breslow and Cain, the disease is assumed rare, the prevalence of factor 1 is .05, and that of factor 2 is .30.

Efficiencies for β_1 and β_2 are based on the model without including an interaction term in the model, while that for β_3, the interaction, is under the assumption $\beta_3 = 0$.

variance estimate used in the latter method involves subtracting a positive term from the naive variance based on logistic regression. Thus the asymptotic relative efficiency of the maximum partial likelihood estimation must be less than 1. However, the point estimates are also different, since the method of Breslow and Cain subtracts an offset based on the actual proportions sampled in various strata, rather than based on the known sampling probabilities. It was thus not clear how the two approaches to analysis would compare for moderate samples.

Accordingly, the operating characteristics of the two approaches were assessed in a small simulation study. Various prevalences are assumed for two dichotomous factors, denoted π_1 and π_2. The two factors again covary with odds ratio θ. Simulations of 1,000 case–control studies were carried out for each of several combinations of parameters and sample sizes, where stochastic frequency matching was followed. In expectation, the total number

of controls equals the total number of cases, but because recruitment is random the actual number of controls recruited varies. N_1 will denote the total number of cases to be studied, and N_0 the (larger) number of controls whose status with regard to factor 1 is to be determined and who are eligible for recruitment. The expected number of recruited (i.e., second-stage) controls is also N_1. When the denominator associated with a particular combination of the factors was 0 a value of .5 was substituted for purposes of analysis.

Table 3 shows the coverage properties of confidence intervals under the two methods. (BC denotes Breslow/Cain and MPLE the unconditional maximum partial likelihood procedure described in Section 3.) For both procedures the empirical bias (not shown) was generally less than 1%. Both methods had coverage consistent with the nominal 95%. As expected, the Breslow/Cain method offers higher power than the unconditional MPLE, but the difference appears generally to be small, especially when the "exposure," the occurrence of factor 1, is rare ($\pi_1 = .05$ in Table 3). These moderate sample results are consistent with the asymptotic relative efficiencies described in Table 2. Results for estimating β_2 are not shown, since the estimates are not different for effects of second-stage variables.

Table 3
Simulation results, based on 1,000 *simulated case–control studies for each combination of parameters*

					Number failing to cover β_1 (95% CI)		Empirical power	
$\exp(\beta_1)$	θ	π_1	N_1	N_0	BC	MPLE	BC	MPLE
2.0	1.0	.3	100	153	52	54	.73	.68
2.0	1.0	.3	150	230	42	40	.90	.84
2.0	1.0	.3	200	307	54	58	.96	.90
2.0	2.0	.3	100	160	59	59	.70	.64
2.0	2.0	.3	150	240	53	57	.88	.83
2.0	2.0	.3	200	321	55	52	.94	.91
3.0	1.0	.3	100	187	51	48	.98	.96
3.0	1.0	.3	150	281	49	51	1.00	1.00
3.0	1.0	.3	200	374	44	50	1.00	1.00
3.0	2.0	.3	100	194	45	48	.98	.95
3.0	2.0	.3	150	291	44	52	1.00	.99
3.0	2.0	.3	200	388	51	45	1.00	1.00
3.0	1.0	.05	100	272	57	60	.74	.72
3.0	1.0	.05	150	409	48	49	.90	.88
3.0	1.0	.05	200	545	50	49	.97	.96
3.0	2.0	.05	100	294	45	48	.77	.75
3.0	2.0	.05	150	442	47	44	.91	.89
3.0	2.0	.05	200	589	43	45	.96	.95

The odds ratio associated with the second factor is assumed to be 2.0. θ denotes the odds ratio for the two factors in the nondiseased population, and π_i denotes the prevalence of factor i in the nondiseased population. π_2 is fixed at .15. The N_0 denote the number of controls to be identified in order to achieve equality on average between numbers of cases and controls.

6. Discussion

The proposed design is appropriate for the commonly encountered situation in case–control studies where certain variables already known to be related to disease risk, such as age and smoking status, are easily obtained. If offers the gains in efficiency associated with matching, but also allows for efficient estimation of effects associated with "matching" factors, and a

flexible analysis that permits the fitting of nonmultiplicative models. There are practical advantages as well. In classical frequency matching, the investigator cannot know how many controls will ultimately be needed to serve as matches in a particular category until case ascertainment is complete. By contrast, the proposed method allows for simultaneous and independent recruitment of cases and controls, and completely avoids the complications associated with filling quotas.

The method does require prior information on screening-variable-specific disease rates or odds ratios. However, if the probability matrix for recruitment was based on erroneous estimates but the actual probabilities used are entered in the offsets, the analysis will still be valid, the only consequence being potentially some loss of efficiency.

While both the computation of asymptotic relative efficiencies and the simulations were done under the assumption that stochastic frequency matching is to be applied, the method described allows for greater flexibility in design, so that distributions of factors can be selected by the investigator to approximately optimize for any selected estimates and comparisons. For example, in the radon study discussed above, one might wish to oversample nonsmoking cases to facilitate characterization of the interaction between smoking and radon. Results in Breslow and Cain suggest great gains in efficiency for estimation of such an interaction can be achieved by approximately equating the numbers of smokers and nonsmokers within each disease group.

If, as in the proposed design, the recruitment decision is made on a subject-by-subject basis, independently (though not identically) among subjects, then there are two choices for analysis. One can use maximum partial likelihood analysis, as described in Section 3, being careful to include the model offset necessary to remove bias in the point estimates. This analysis offers the advantage of maximization of the likelihood by means of logistic regression in packaged programs like GLIM, with simple variance estimates and valid likelihood ratio testing. The likelihood is based, however, on partial information, since the original recruitable sample identified is not fully exploited and thus some information related to the screening variables is lost. If we instead apply the method proposed by Breslow and Cain, all of the available data are used, but the variance estimation is more complex and the method does not provide for likelihood ratio testing.

It is important to note that the maximum partial likelihood analysis is appropriate only when the subsampled individuals are selected by a Bernoulli sampling mechanism. If the investigator has simply selected subsamples of a convenient size, then the method of Breslow and Cain should be used. Thus, for example, in the stratified sampling example described by Fears and Brown (1986; see also Breslow and Zhao (1988)), the methods of Section 3 should not be applied. Similarly, if traditional frequency or quota matching has been followed, then one can still estimate effects of the matching factors, but only by applying the method of Breslow and Cain.

Following a design employing biased, individually randomized recruitment, the choice of analysis will depend ultimately on which factor is of primary interest in the study. If, as in Breslow and Cain, the exposures of interest are among the easily ascertained first-stage data, and the data obtained after randomization to recruitment are needed only to adjust the estimated effect of interest, then one would prefer the slightly more efficient method of Breslow and Cain. If the variable of primary interest is obtained at the second stage, then the maximum partial likelihood method might be preferred, because it offers the same efficiency for the parameter of primary interest but provides for flexible exploratory analyses by means of likelihood ratio testing.

Acknowledgements

The authors wish to thank Drs N. Breslow, C. Brown, B. Gladen, T. Fears, M. Gail, J. Lubin, D. Sandler, C. Shy, and E. White for their helpful comments.

Résumé

Un plan d'étude cas–témoin est proposé dans lequel la sélection des sujets pour lesquels on recueille toutes les variables est basée à la fois sur le statut cas/témoin et sur des variables de sélection faciles à obtenir qui peuvent être liées à la maladie. Le recrutement des sujets se poursuit au fur et à mesure que les cas sont diagnostiqués et se fait selon un échantillonnage de Bernoulli indépendant, avec des probabilités de recrutement fixées au préalable par l'investigateur. En particulier, le processus d'échantillonnage peut être conçu pour assurer en moyenne un appariement sur les distributions marginales à condition, comme c'est souvent le cas, d'avoir des estimations à priori des taux de la maladie ou des odds-ratios associés aux variables de sélection telles que l'âge et le sexe. Par ailleurs, lorsqu'on étudie une exposition rare, un recrutement biaisé peut permettre d'enrichir l'échantillon avec certains groupes de sujets. Avec un tel plan d'étude il y a deux analyses possible basées sur la régression logistique, qui permettent toutes les deux d'estimer les effets associés aux variables de sélection utilisées pour biaiser le recrutement.

References

Baker, R. J. and Nelder, J. A. (1978). *The GLIM System: Release 3*. Oxford: Numerical Algorithms Group.

Breslow, N. E. and Cain, K. C. (1988). Logistic regression for two-stage case–control data. *Biometrika* **75**, 11–20.

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research I: The Analysis of Case–Control Studies*. Lyon: International Agency for Research on Cancer.

Breslow, N. E., Day, N. E., Halvorsen, K. T., Prentice, R. L., and Sabai, C. (1978). Estimation of multiple relative risk functions in matched case–control studies. *American Journal of Epidemiology* **108**, 299–307.

Breslow, N. E. and Zhao, L. P. (1988). Logistic regression for stratified case–control studies. *Biometrics* **44**, 891–899.

Cain, K. C. and Breslow, N. E. (1988). Logistic regression analysis and efficient design for two-stage studies. *American Journal of Epidemiology* **128**, 1198–1206.

Cohen, B. L. (1987). Tests of the linear, no-threshold dose–response relationship for high-LET radiation. *Health Physics* **52**, 629–634.

Fears, T. R. and Brown, C. C. (1986). Logistic regression methods for retrospective case–control studies using complex sampling procedures. *Biometrics* **42**, 955–960.

Gail, M. H. (1988). The effect of pooling across strata in perfectly balanced studies. *Biometrics* **44**, 151–162.

Kleinbaum, D. G., Kupper, L. L., and Morgenstern, H. (1982). *Epidemiologic Research: Principles and Quantitative Methods*. Belmont, California: Lifetime Learning Publications.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case–control studies. *Biometrika* **66**, 403–411.

Thomas, D. C. and Greenland, S. (1985). The efficiency of matching in case–control studies of risk-factor interactions. *Journal of Chronic Diseases* **38**, 569–574.

Walker, A. M. (1982). Anamorphic analysis: Sampling and estimation for covariate effects when both exposure and disease are known. *Biometrics* **38**, 1025–1032.

White, J. E. (1982). A two-stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology* **115**, 119–128.

Woolson, R. F., Bean, J. A., and Rojas, P. B. (1986). Sample size for case–control studies using Cochran's statistic. *Biometrics* **42**, 927–932.

Received January 1989; *revised July and September* 1989; *accepted October* 1989.

Appendix

Consider what happens when conditional logistic regression is to be carried out. Biased recruitment implies that the distribution $\Pr(\mathbf{X} \mid D = i)$ has been distorted in a systematic way. We can write

$$\Pr[\mathbf{X} \mid R = 1, D = i] = \Pr[\mathbf{X} \mid D = i](\Pr[R = 1 \mid D = i, \mathbf{X}]/\Pr[R = 1 \mid D = i]). \quad (A.1)$$

Under the usual assumptions sampling depends on disease status but not on other variables, so that the ratio factor to the right becomes 1. However, following biased recruitment it must be included in the likelihood.

Following the development described heuristically in Breslow and Day (1980, pp. 204–205), and worked out rigorously by Breslow et al. (1978), if a stratum contains n_1 cases and n_0 controls, where $n = n_1 + n_0$, then if we condition on the n variable vectors observed, $\mathbf{X}_1, \ldots, \mathbf{X}_n$, the conditional probability that the first n_1 vectors belonged to the cases can be written

$$\frac{\prod_{j=1}^{n_1} \Pr[\mathbf{X}_j \mid R = 1, D = 1] \prod_{j=n_1+1}^{n} \Pr[\mathbf{X}_j \mid R = 1, D = 0]}{\sum_{\pi \in S} \prod_{j=1}^{n_1} \Pr[\mathbf{X}_{\pi(j)} \mid R = 1, D = 1] \prod_{j=n_1+1}^{n} \Pr[\mathbf{X}_{\pi(j)} \mid R = 1, D = 0]},$$

where S is the set of all permutations on the set $\{1, 2, \ldots, n\}$. Now applying (A.1), and using the fact that $\Pr[\mathbf{X} \mid D = i] = \Pr[D = i \mid \mathbf{X}]\Pr[\mathbf{X}]/\Pr[D = i]$, we can rewrite this as

$$\frac{\prod_{j=1}^{n_1} \Pr[D = 1 \mid \mathbf{X}_j]\Pr[R = 1 \mid D = 1, \mathbf{X}_j] \prod_{j=n_1+1}^{n} \Pr[D = 0 \mid \mathbf{X}_j]\Pr[R = 1 \mid D = 0, \mathbf{X}_j]}{\sum_{\pi \in S} \prod_{j=1}^{n_1} \Pr[D = 1 \mid \mathbf{X}_{\pi(j)}]\Pr[R = 1 \mid D = 1, \mathbf{X}_{\pi(j)}] \prod_{j=n_1+1}^{n} \Pr[D = 0 \mid \mathbf{X}_{\pi(j)}]\Pr[R = 1 \mid D = 0, \mathbf{X}_{\pi(j)}]}.$$

We can apply the factorization of recruitment probability developed above to rewrite this as

$$\frac{\prod_{j=1}^{n_1} \Pr[D = 1 \mid \mathbf{X}_j]\Pr[R = 1 \mid I, D = 1, \mathbf{X}_j] \prod_{j=n_1+1}^{n} \Pr[D = 0 \mid \mathbf{X}_j]\Pr[R = 1 \mid I, D = 0, \mathbf{X}_j]}{\sum_{\pi \in S} \prod_{j=1}^{n_1} \Pr[D = 1 \mid \mathbf{X}_{\pi(j)}]\Pr[R = 1 \mid I, D = 1, \mathbf{X}_{\pi(j)}] \prod_{j=n_1+1}^{n} \Pr[D = 0 \mid \mathbf{X}_{\pi(j)}]\Pr[R = 1 \mid I, D = 0, \mathbf{X}_{\pi(j)}]}.$$

Dividing the numerator and denominator by the product across all subjects of

$$\Pr[R = 1 \mid I, D = 0, \mathbf{X}_j],$$

we obtain the following simplification:

$$\frac{\prod_{j=1}^{n_1} \Pr[D = 1 \mid \mathbf{X}_j] \dfrac{\Pr[R = 1 \mid I, D = 1, \mathbf{X}_j]}{\Pr[R = 1 \mid I, D = 0, \mathbf{X}_j]} \prod_{j=n_1+1}^{n} \Pr[D = 0 \mid \mathbf{X}_j]}{\sum_{\pi \in S} \prod_{j=1}^{n_1} \Pr[D = 1 \mid \mathbf{X}_{\pi(j)}] \dfrac{\Pr[R = 1 \mid I, D = 1, \mathbf{X}_{\pi(j)}]}{\Pr[R = 1 \mid I, D = 0, \mathbf{X}_{\pi(j)}]} \prod_{j=n_1+1}^{n} \Pr[D = 0 \mid \mathbf{X}_{\pi(j)}]},$$

which is simply

$$\frac{\prod_{j=1}^{n_1} \exp(\mathbf{X}_j(\boldsymbol{\beta} + \boldsymbol{\gamma}))}{\sum_{\pi \in S} \prod_{j=1}^{n_1} \exp(\mathbf{X}_{\pi(j)}(\boldsymbol{\beta} + \boldsymbol{\gamma}))}.$$